ANSTO/E-776

Australian Government

Nuclear-based science benefiting all Australians

Receptor Modelling using PMF2: influence on source fingerprints and their contributions using different input options and measurement uncertainties

by

Jagoda Crawford

David D. Cohen

Prepared within the Institute for Environmental Research
Australian Nuclear Science and Technology Organisation

November 2012

Abstract

Receptor modelling is used to determine source fingerprints, from measurements at the receptor site, and then apportion the measured mass to the identified possible sources. Mathematical techniques used to solve this problem include Principal Component Analysis and multivariate techniques such as Positive Matrix Factorisation and UNMIX.

In this report PMF analysis is carried out using the PMF2 program. When using PMF2 a number of options control the optimisation process and by varying the options the final solution might be impacted. In addition the uncertainties associated with the measurement have an impact on the solution. Here we systematically vary the PMF2 options and the specified uncertainties of each measurement and compare each solution to a base case solution which has recently been published for the selected dataset.

The main findings were that the uncertainties in the measurements and the number of factors chosen to be recovered were the key quantities for the PMF analysis of the considered dataset.

Keywords: PMF, receptor modelling

ISSN 10307745
ISBN 1 921268 17 4

Table of Contents

List of Figures

List of Tables

1 Introduction

Receptor modelling is used to determine source fingerprints, from measurements at the receptor site, and then apportion the measured mass to the identified possible sources (Hopke 1991, Henry 2002). Mathematical techniques used to solve this problem include Principal Component Analysis (PCA Jollife, 1989), and multivariate techniques such as Positive Matrix Factorisation (PMF; Paatero and Tapper, 1994) and UNMIX (Henry, 2002).

In this report PMF analysis is carried out using the PMF2 program (Paatero, 2010). When using PMF2 a number of options control the optimisation process and by varying the options the final solution might be impacted. In addition the errors associated with the measurement have an impact on the solution. Here we systematically vary the PMF2 options and the specified error of each measurement and compare each solution to a *base case* solution which has recently been published for the selected dataset (Cohen et al., 2012).

2 Study site and aerosol sampling

2.1 Study site and local meteorology

Aerosol samples have been collected at Richmond in New South Wales (NSW) Australia from January 2001. Richmond is within the Sydney basin which is bounded by high altitude land to the south, west and north, with Richmond located on the western edge of the basin.

At the site, prevailing wind direction in the summer is onshore winds, also resulting in higher precipitation. The cooler half of the year is associated with low wind speeds and low precipitation conditions from westerly wind during the night and morning period.

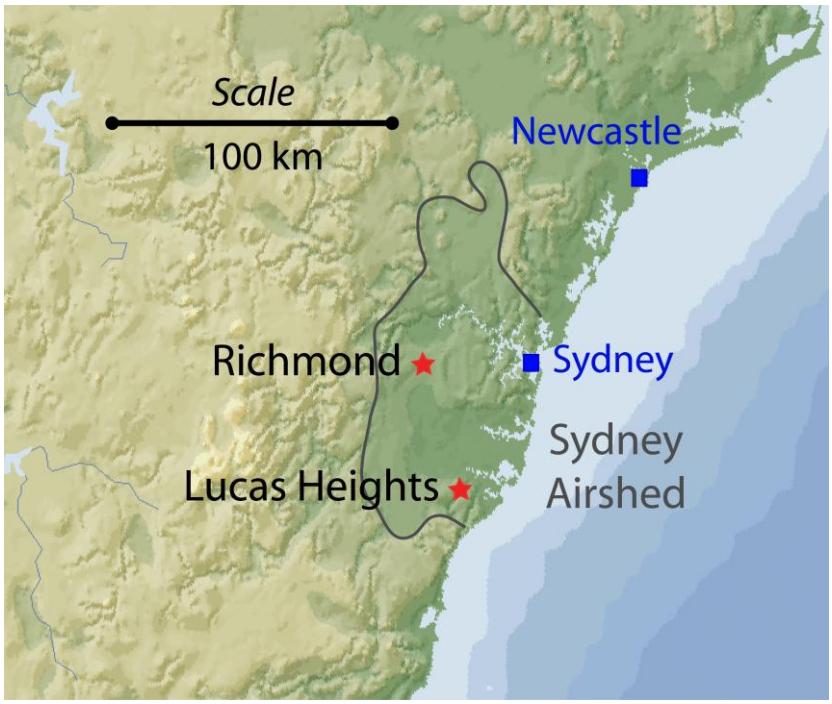


Figure 1: The Richmond Study Site located in the Sydney Airshed.

2.2 Aerosol sampling and elemental analysis

The aerosol sampling program constituted 24-hour integrated samples (from midnight to midnight local time) taken twice a week (Wednesday and Sunday) using a cyclone $PM_{2.5}$ system with 22 l/min flow rate (Cohen et al., 1996). While measurements continue to be made at the site, for this study measurements up to December 2011 have been used.

Accelerator-based ion beam analysis (IBA) techniques were used to perform the elemental analyses of the aerosol samples (Cohen, 1998; Cohen et al. 2004). These techniques can provide quantitative elemental information on a broad range of elements including: H, Na, Al, Si, P, S, Cl, K, Ca, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Br, Pb. Laser absorption methods were used to determine the black carbon (BC) concentrations (Cohen et al., 2000, and Taha et al., 2007).

3 PMF problem formulation

The basic PMF problem can be specified as (Paatero, 2010):

$$X = GF + E \qquad (1)$$

or

$$x_{i,j} = \sum_{k=1}^{p} g_{i,k} f_{k,j} + e_{i,j} \qquad (2)$$

where the matrix X contains the measured quantities, i.e. $x_{i,j}$ represents the concentration of chemical species j in the i^{th} sample. Matrices G and F are factor matrices to be determined and E is the matrix of residuals. If n observations are available, each containing m chemical species and if a p-factor model is being considered, G is an n *by* p matrix of source contributions, describing the temporal variation of the source strength. The matrix F is a p by m matrix of source chemical compositions, or source fingerprints.

The objective of positive matrix factorisation (PMF) is to minimise the penalty function Q under the constraints that the factor elements remain non-negative:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{e_{i,j}^2}{s_{i,j}^2} \qquad (3)$$

where $s_{i,,j}$ is a specified error for each data value. The following form of $s_{i,j}$ is used:

$$s_{i,j} = MDL_{i,j} + Error_{i,j} \max(|x_{i,j}|, |y_{i,j}|) \qquad (4)$$

where, $MDL_{i,j}$ is the specified minimum detectible limit and $Error_{i,j}$ is the specified statistical error to account for experimental error, peak area determination and calibration, and $y_{i,j}$ is the fitted value i.e. Y=GF. The theoretical Q value is determined from equation 5, being the degrees of freedom of the problem.

$$Q_{thry} = nm - p(n+m) \qquad (5)$$

For a well-defined problem the ratio of the minimum Q from the optimisation (i,e, the experimental Q value), to Q_{thry} ($R_Q = Q_{expt}/Q_{thry}$) is expected to be close to 1.

4 Base Case

The *base case* consists of 912 samples for which the concentration of 21 elements were determined. The mean, median and standard deviation of the elemental composition of these samples is presented in Table 1. Also presented in Table 1 are the mean, median and standard deviation of the minimum detection limit (MDL) associated with each element in each sample and also that of the statistical error associated with each element of each sample.

Table 1: Sample mean, median and standard deviation of the elemental composition, the MDL and the standard deviation associated with each element.

Element	Data (ng/m^3)			MDL (ng/m^3)			Statistical Error		
	Mean	Median	Std	Mean	Median	Std	Mean	Median	Std
H	252.48	195.50	201.85	11.91	7.8	8.48	0.07	0.06	0.02
Na	181.11	97.95	254.13	156.66	164.0	51.68	0.33	0.35	0.14
Al	19.49	8.90	40.54	5.75	5.7	2.18	0.41	0.26	0.62
Si	69.50	40.45	113.00	2.93	2.9	0.97	0.08	0.07	0.03
P	6.73	3.90	8.55	2.98	3.0	1.05	0.50	0.31	0.73
Si	399.55	299.20	317.86	2.51	2.5	0.84	0.06	0.06	0.00
Cl	131.98	39.70	213.60	2.75	2.7	0.90	0.23	0.07	0.58
K	58.30	40.05	55.56	1.46	1.5	0.41	0.06	0.06	0.01
Ca	16.18	12.90	12.40	1.67	1.6	0.56	0.10	0.08	0.07
Ti	2.59	1.70	3.65	1.10	1.1	0.34	0.33	0.28	0.20
V	0.31	0.20	0.34	0.98	1.0	0.29	0.91	0.85	0.62
Cr	0.47	0.35	0.88	0.77	0.8	0.25	0.71	0.59	0.42
Mn	1.13	1.00	0.86	0.72	0.7	0.26	0.39	0.34	0.20
Fe	27.39	21.10	29.60	0.71	0.7	0.26	0.07	0.07	0.01
Co	0.33	0.20	0.59	1.42	1.5	0.53	1.28	1.17	1.06
Ni	0.38	0.20	1.11	0.78	0.8	0.29	0.81	0.75	0.54
Cu	1.55	1.00	1.97	0.89	0.9	0.36	0.47	0.40	0.38
Zn	6.14	4.30	6.67	1.07	1.1	0.44	0.23	0.18	0.16
Br	2.43	2.00	1.98	3.45	3.7	1.21	0.78	0.57	0.87
Pb	4.89	3.20	5.59	6.61	7.0	2.33	0.78	0.61	0.82
BC	826.22	710.05	465.69	28.53	28.6	0.80	0.08	0.08	0.00

For this dataset seven source fingerprints have been resolved (Cohen et al., 2012), which forms the *base case* in this study. The experimental Q value (Q_{expr}) was 6,709 as opposed to Q_{thry} of 12,621. Thus $R_Q=Q_{expt}/Q_{thry}$ = 0.53. Based on the key elements in each factor, these factors were associated with the following possible sources (Cohen et al., 2012):

2ndryS: Secondary sulfate fingerprint, produced by the conversion of SO_2 gas to the particulate sulfate phase in the presence of sunlight and water vapour (Seinfeld and Pandis, 1998). The ratio of S/H was 3.46 indicating that the majority of it was fully neutralised ammonium sulfate $(NH_4)_2SO_4$.

Soil: This factor contains the five key elements (Al, Si, Ca, Ti and Fe) commonly associated with windblown soils. The Al/Si ratio of 0.325 was obtained. Typically the Al/Si ratio for soils is between (0.25-0.35).

Sea: This factor is dominated by Na and Cl, with small amounts of Br which is indicative of fresh sea spray. The obtained Cl/Na ratio was 2.38 which was higher than the theoretical of 1.54 for NaCl.

IndSaged: This factor is dominated by Na, S and BC. The Na to S ratio was indicative of Na_2SO_4 probably formed by chemical reactions of sea spray particles with sulfate particles. This forms the aged sea salt source i.e. after reaction with industrial S.

Smoke: This factor was dominated by high H from organic sources, K and BC from biomass burning.

Industry This factor was dominated by BC, Fe and Zn.

Auto: This factor was dominated by H, BC and trace elements associated with motor vehicles such as Zn from tyre wear, P and Ca from engine oils and small amounts of Pb and Br associated with historic leaded petrol use.

Scatter plots of key elements of *2ndryS*, *Soil* and *Sea* are presented in Figure 2. The ratio of S/H for *2ndryS* was 3.46 (represented by the black line in Figure 2a) which was close to the theoretical ratio of 4 for $(NH_4)_2SO_4$ (represented in green). These two lines are close to each other on one edge of the scatter plot of S against H. The Cl/Na ratio in the *Sea* fingerprint was 2.38 which was on an edge in the scatter plot of Cl against Na (black line in Figure 2c) and the corresponding theoretical ratio of 1.54 for NaCl is represented in green. The Al/Si ratio in the *Soil* fingerprint was 0.325 (black line in Figure 2b) which is between 0.25 and 0.35 (green lines).

The metrics used in the comparison of solutions under different options with the *base case* solution include R_Q, Cl/Na for the *Sea* fingerprint, S/H for the *2ndryS* fingerprint, Al/Si for the *Soil* fingerprint and the percentage contribution to the total mass from each of the identified fingerprints.

5 Selecting the number of source fingerprints

5.1 PMF analysis using the full dataset

As a first step a PMF analysis was undertaken for 6, 7 and 8 source factors, after removing one extreme event from the data, i.e. that of a severe dust storm on 20/09/2009. The factors associated with *IndSaged* and *2ndryS* (for the seven factor solution) were combined into one factor (for the six factor solution). In addition, when six factors were resolved, some sulfur was allocated to the factors associated with *Smoke* and *Auto* (Figure 3) and BC was allocated to the factors associated with *Sea* and *Industry*, which was not present when seven factors were resolved. On the other hand when eight factors were resolved a factor predominantly with H and S was identified (*H.S*). The column in the G matrix corresponding to factor *H.S* was correlated with the column corresponding to *Smoke* (Figure 4), indicating the possibility that smoke emissions were split into two factors.

The Q values are presented in Table 2, and so is the S/H ratio for the *2ndryS* factor, Al/Si ratio for *Soil* and Cl/Na ratio for *Sea*. The ratio R_Q for the eight factors is quite small, indicating that eight factors possibly do not represent the optimum factorisation. The Al/Si ratio indicates no change in the *Soil* fingerprint; however, there is a large change in the S/H ratio for *2ndryS*. When comparing these rations one needs to keep in mind that for the six factor case the *2ndryS* fingerprint is a combination of *2ndryS* and *IndSaged*.

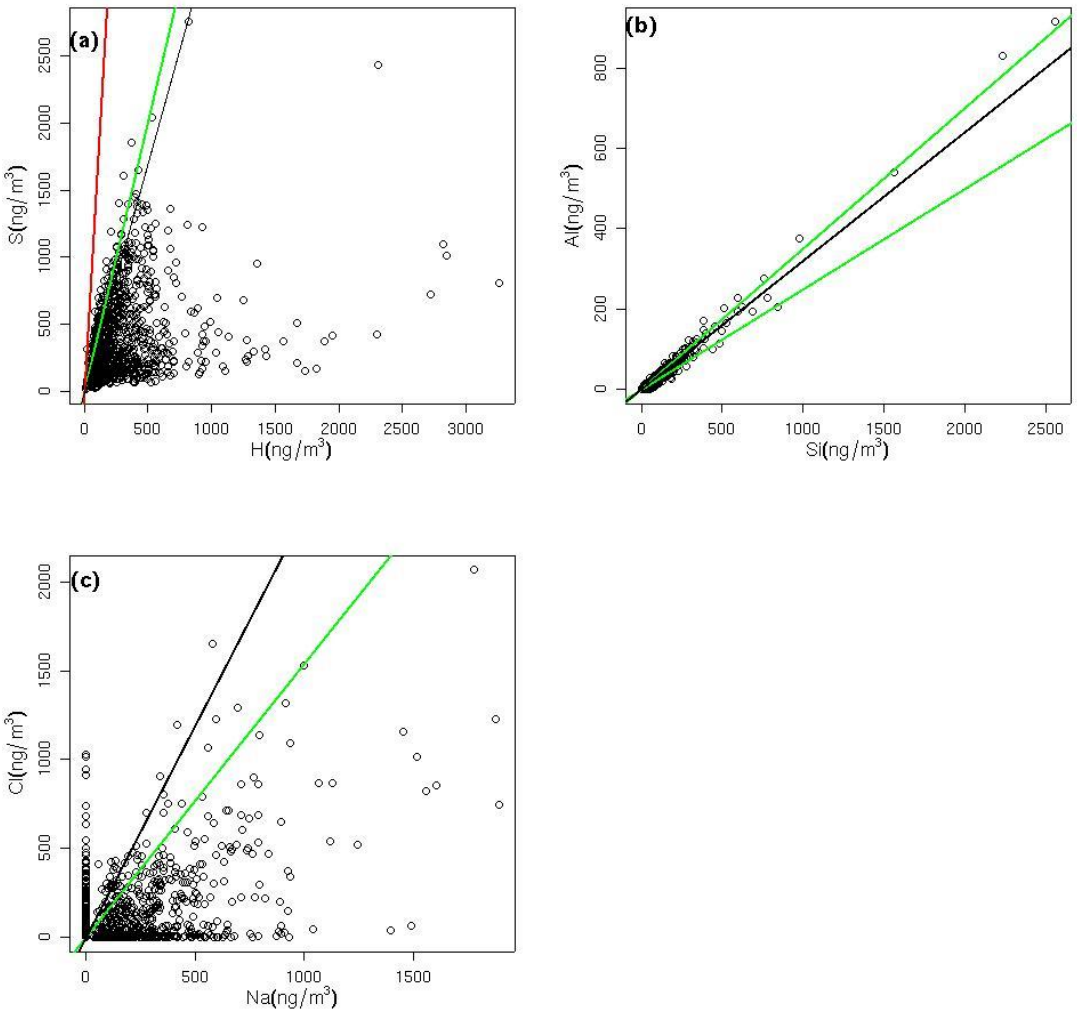
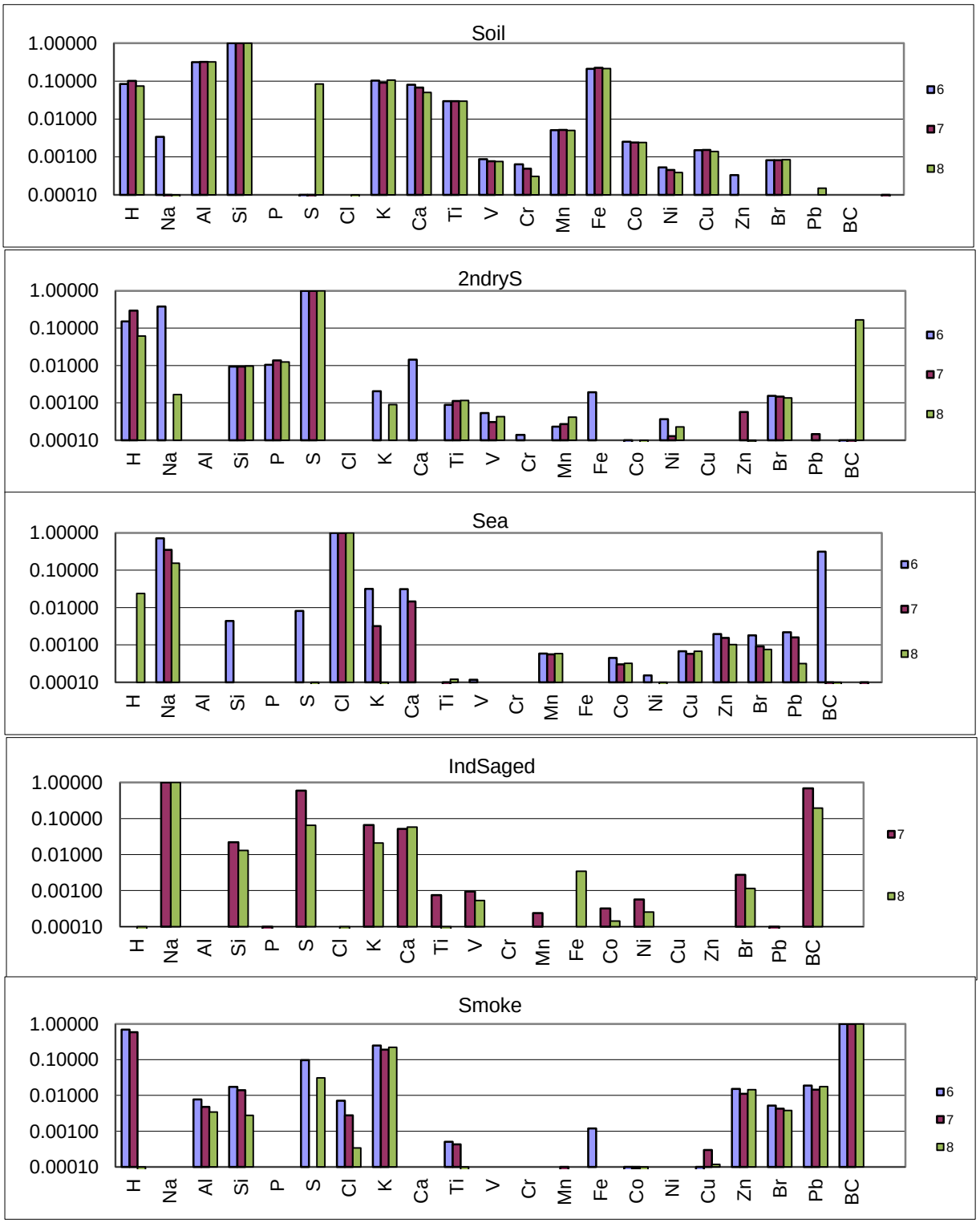


Figure 2: Plots of (a) S against H with the black line indicating the ratio obtained for *2ndryS* fingerprint, for the seven factor solution, the green line indicating the theoretical ratio of 4 for $(NH_4)_2SO_4$, and the red line indicating the theoretical ratio of 16 for of H_2SO_4, (b) Al against Si with the black line indicating the ratio obtained for *Soil* fingerprint, the green lines indicating ratios of 0.25 and 0.35, (c) Cl against Na with the black line indicating the ratio obtained in *Sea* for the seven factor solution and the green line indicating the theoretical value of 1.54 for NaCl.

Table 2: PMF2: Q values for different number of factors and the ratios of key elements for some of the fingerprints.

Factors	Q_{thry}	Q_{expt}	Q_{expt}/Q_{thry}	S/H in 2ndryS	Al/Si in Soil	Cl/Na in Sea
6	15381	12466	0.8	6.64	0.32	1.40
7	14329	8324	0.6	3.40	0.32	2.87
8	13277	5409	0.4	16.32	0.32	6.46



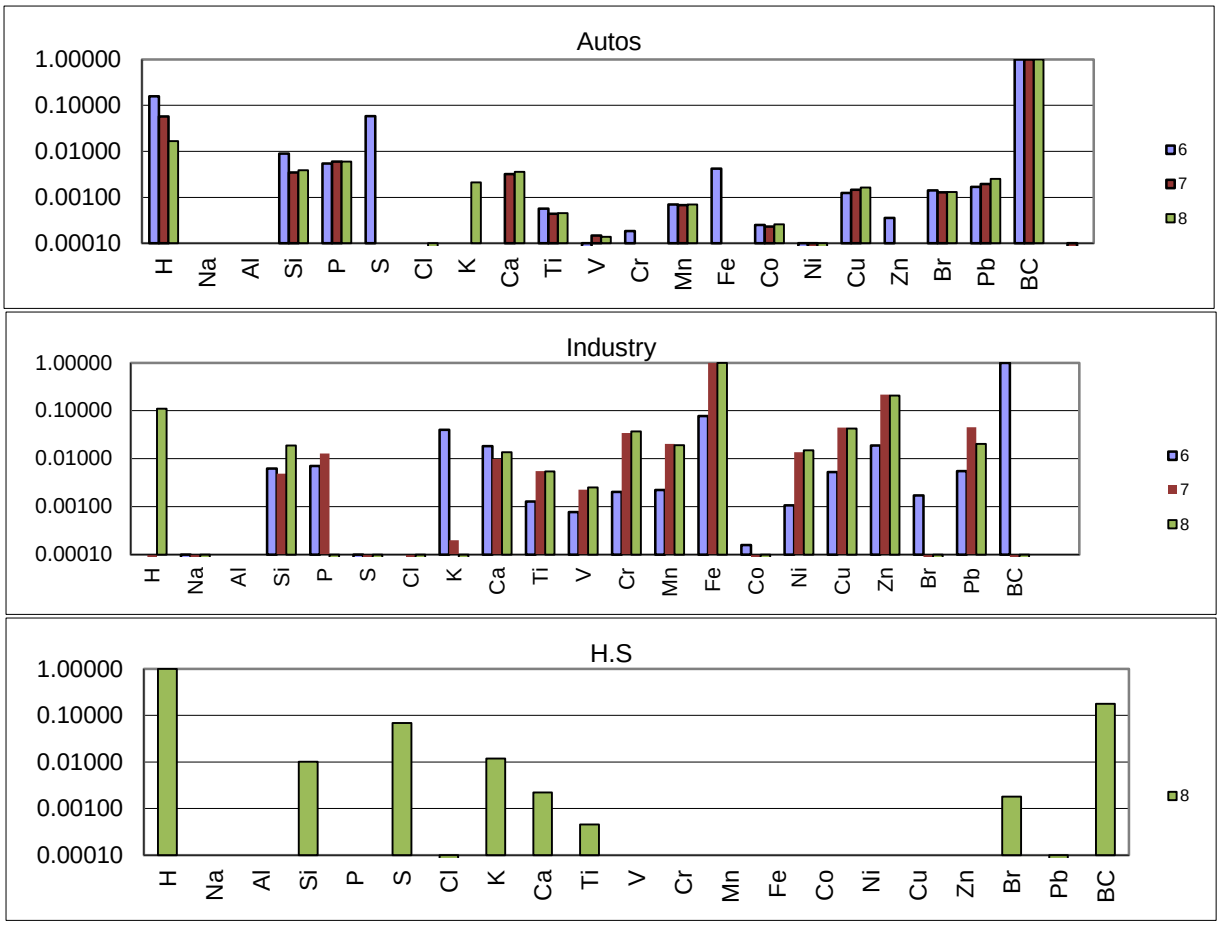


Figure 3: Source fingerprints when solving for 6 (blue), 7 (brown) and 8 (green) factors using the complete set of data, with only one outlier removed, that on 20/9/2009. Within each fingerprint elemental ratios are normalised, using the PMF2 option MaxF=1 set to T.

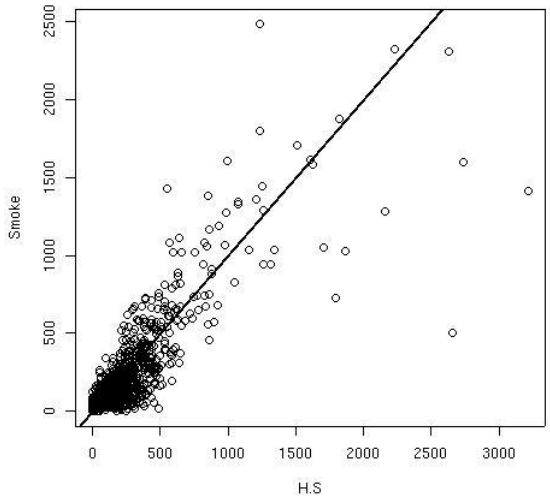


Figure 4: Correlation between column of the G matrix for *Smoke* and *H.S* factor.

The PMF fitted $PM_{2.5}$ mass concentrations, when seven factors were resolved, are presented against the measured $PM_{2.5}$ mass concentrations in Figure 5. From Figure 5 we can see that all of the measurements were not reproduced by the seven factor

factorisation. The next step was to remove outliers from the dataset and perform the PMF analysis on the reduced dataset.

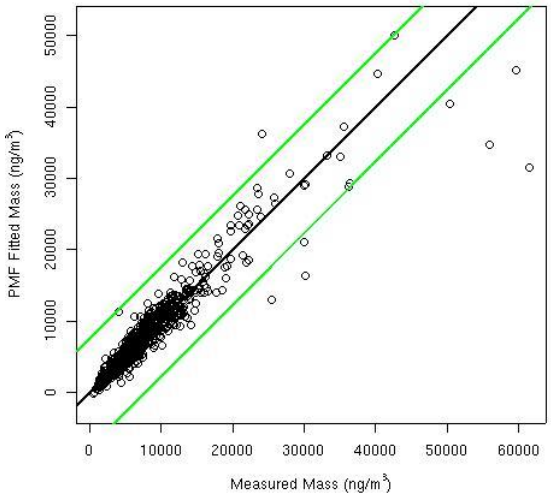
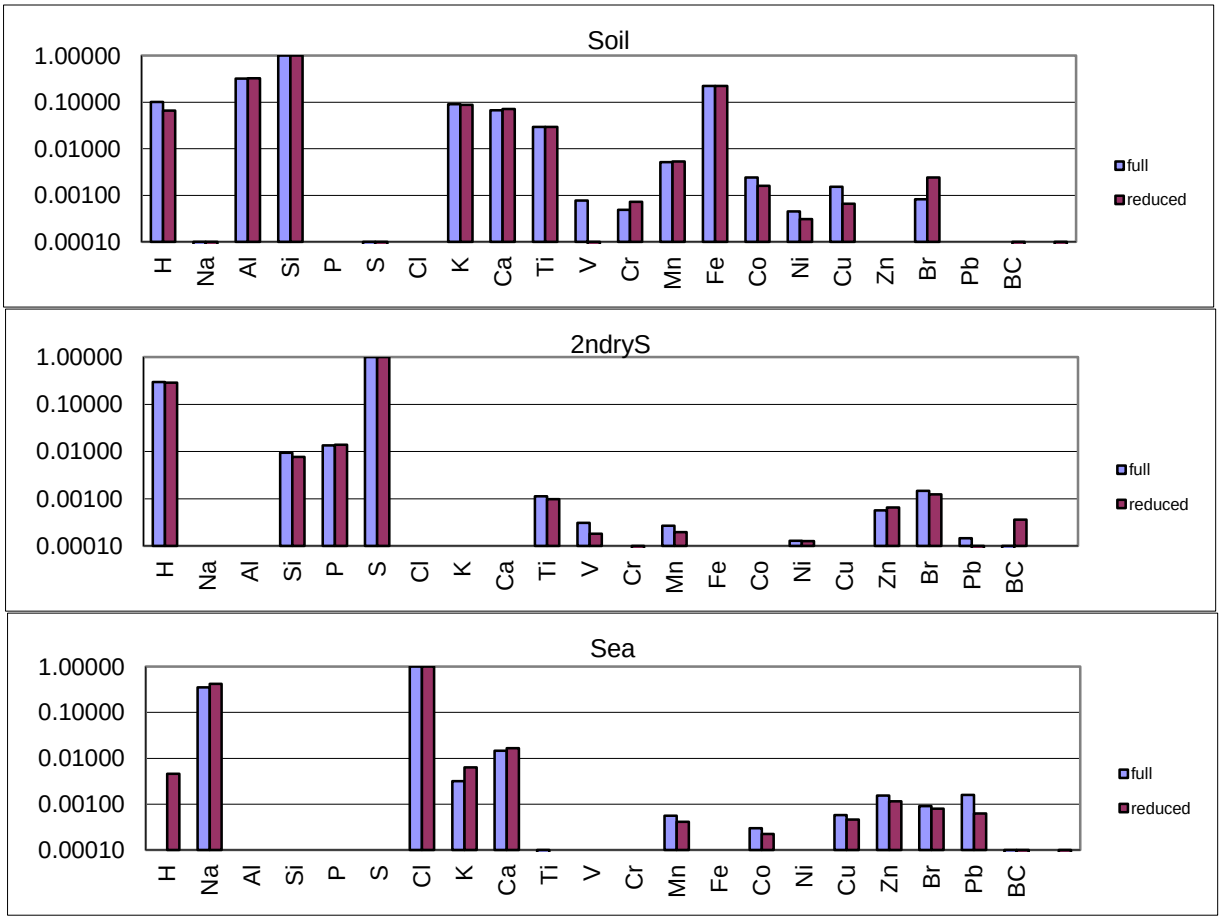


Figure 5: PMF fitted mass concentrations against measured mass concentrations when seven factors were resolved. Points between green lines are less than 4 standard deviations away from the black line.

5.2 PMF analysis using the reduced dataset

After removing those samples which resulted in outlier values in Figure 5, the seven factor PMF solution is presented in Figure 6 (in brown while the corresponding seven factors for the full dataset is presented in blue). There are some small differences in the elemental composition of the factors with the major one being the presence of BC in the *Industy* factor. The scatter plot of the fitted $PM_{2.5}$ mass concentrations against the measured mass concentrations is presented in Figure 7, which shows no outliers.



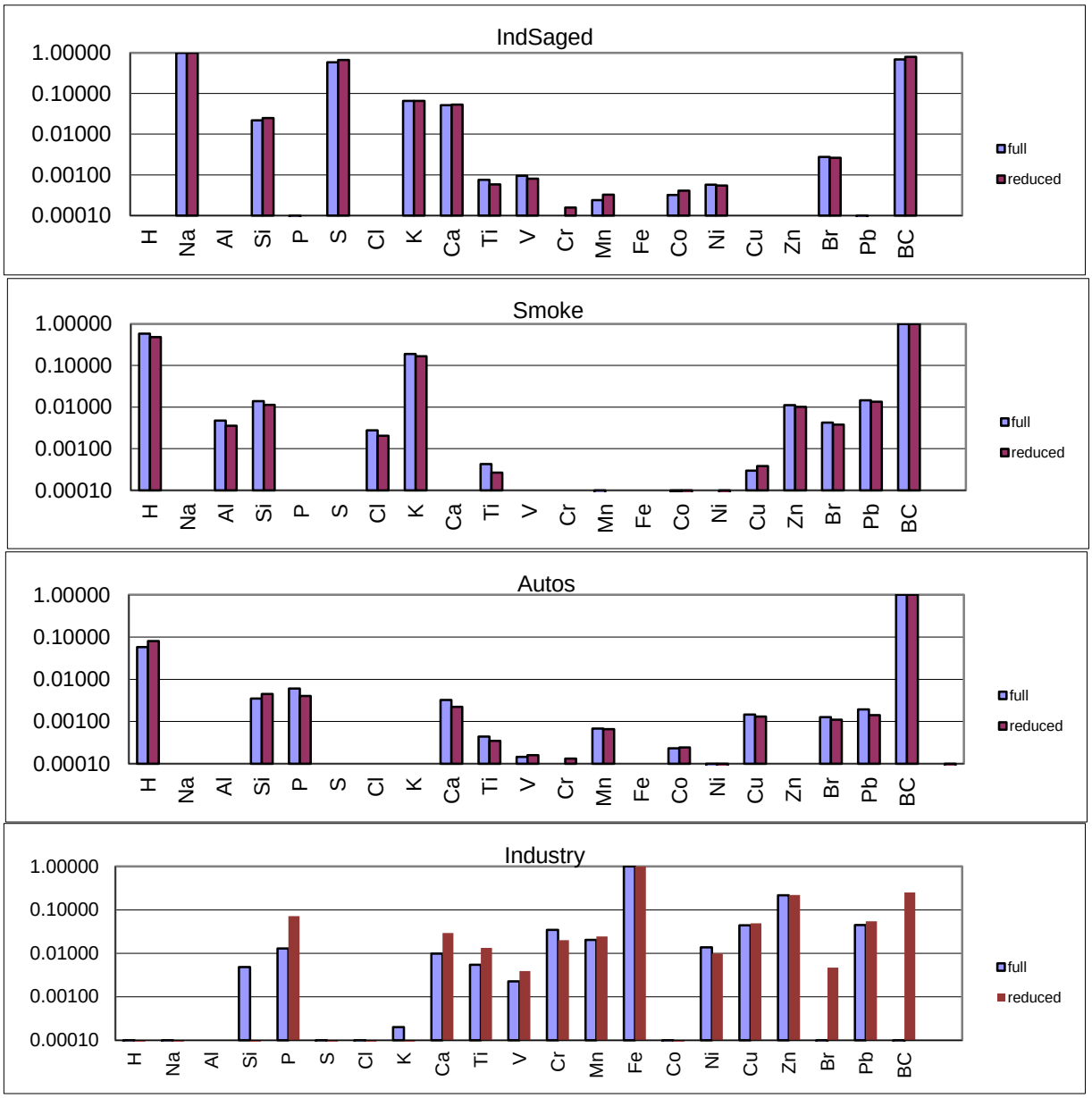


Figure 6: Fingerprints (normalised using MaxF=1) for the seven factor PMF solution; for the full dataset in blue and the reduced dataset in brown.

The reduced dataset and the associated fingerprints form the *base case* against which all later solutions were tested.

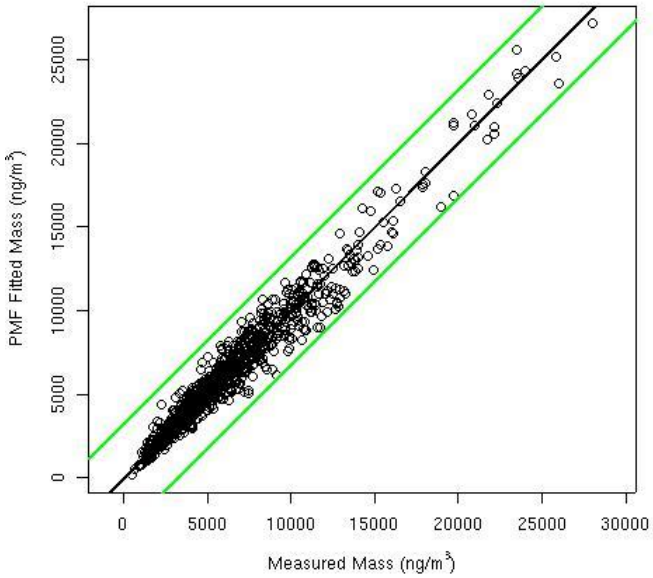


Figure 7: PMF fitted mass concentrations against measured mass concentrations for seven factors when the reduced dataset is used. All points are within 4 standard deviations from the black line.

6 Normalisation of the F matrix

Normalising the F matrix by selecting either the sum of the rows to equal to one (Sum|F| = 1) or the maximum element in each row to equal to one (MaxF=1) resulted in no difference in the solution or the Q_{expt} value.

7 Effects of the iteration control parameters

In this section the results of varying the *lims* and χ^2-test parameters for the three levels of iteration are considered (i.e how close factor elements can get to zero and how much change in the Q_{expt} value is acceptable for convergence, i.e. the χ^2 value represents the relative change in Q_{expt} between two successive iterations). The *base case* values used are specified in Table 3. In the following sections the results when varying each of these parameters in turn, while keeping all others to the *base case* value, are presented.

Table 3: Iteration control parameters.

Level	*lims*	χ^2 test
1	10.00	0.50
2	0.30	0.010
3	0.000005	0.000001

7.1 Changes in lims

7.1.1 Level 1 *lims*

Three values were considered for the level 1 *lims* parameter, 10 (*base case*), 1 and 0.1. There was little difference in the final solution, as seen by the factor contributions

and the partitioning of the total mass to each of the identified source fingerprints (Table 4). The main difference was in the number of iterations to convergence (Table 5), which shows that the smaller the value of level 1 *lims* the less iteration to final convergence. The S/H ratio for the *2ndryS* fingerprint, the Al/Si ratio for the *Soil* fingerprint and the Cl/Na ratio for *Sea* fingerprint are presented in Table 6, where no difference is seen for the level 1 *lims* parameter.

Table 4: Percentage contribution to the total mas from each of the identified fingerprints, when the different *lims* values are used.

	Base Case	Level 1		Level 2		Level 3	
lims value		1	0.1	0.03	0.003	0.0005	0.005
2ndryS	25.18	25.18	25.17	25.28	25.15	25.19	25.28
Soil	5.59	5.59	5.59	5.59	5.59	5.59	5.59
Sea	5.17	5.18	5.15	5.12	5.13	5.18	5.18
IndSaged	14.46	14.46	14.46	14.45	14.49	14.45	14.38
Smoke	33.77	33.75	33.69	33.74	33.76	33.74	33.70
Autos	14.51	14.52	14.56	14.43	14.53	14.51	14.44
Ind.	1.32	1.32	1.38	1.39	1.35	1.34	1.43

Table 5: Number of iterations for each *lims* option and the final Q_{expt} values.

Level 1 lims			Level 2 lims			Level 3 lims		
Value	Iterations	Q_{expt}	Value	Iterations	Q_{expt}	Value	Iterations	Q_{expt}
1.00	170	6709	0.030	182	6699	0.000500	166	6709
0.10	159	6708	0.003	218	6699	0.005000	155	6712
Base Case								
10.00	174	6709	0.300	174	6709	0.000005	174	6709

Table 6: The S/H ratio for *2ndyS* fingerprint, Al/Si ratio for the *Soil* fingerprint and Cl/Na for the *Sea* fingerprint for the various values of the *lims* parameter, compare to S/H ratio of 3.46, Al/Si ratio of 0.325 and Cl/Na ratio of 2.38 for the base case.

Level 1 lims				Level 2 lims				Level 3 lims			
Value	S/H	Al/Si	Cl/Na	Value	S/H	Al/Si	Cl/Na	Value	S/H	Al/Si	Cl/Na
1.00	3.46	0.325	2.38	0.030	3.42	0.325	2.39	0.0005	3.46	0.326	2.38
0.10	3.46	0.325	2.38	0.003	3.44	0.325	2.39	0.0050	3.45	0.325	2.39

7.1.2 Level 2 *lims*

Three values were considered for the level 2 *lims* parameter, 0.3 (*base case*), 0.03 and 0.003. There was little difference in the final solution, as seen by the factor contributions and the partitioning of the total mass to each of the identified source fingerprints (Table 4). The main difference in the F matrix factors was an increase of BC in the *2ndyS* fingerprint from 0.00036 to 0.01 and a decrease of BC in the *Industry* fingerprint. In terms of iterations to convergence, the smaller the value of the level 2 *lims* the more iteration were required to final convergence (Table 5).

7.1.3 Level 3 *lims*

Three values were considered for the level 3 *lims* parameter, 0.005, 0.0005 and 0.000005(*base case*). The main difference was in the number of iterations required for convergence (Table 5) and the F factor composition of the trace elements. This is not surprising as this value controls the final amount of pull to zero of factor entries. A comparison of the *base case* F factors and those when a value of 0.005 is used for *lims* level 3 is presented in Figure 8.

7.2 Changes in χ^2

Changing the level 1, level 2 and level 3 χ^2 had very little impact as supported by the results in Tables 7, 8 and 9.

Table 7: Number of iterations for each χ^2 option and the final Q values.

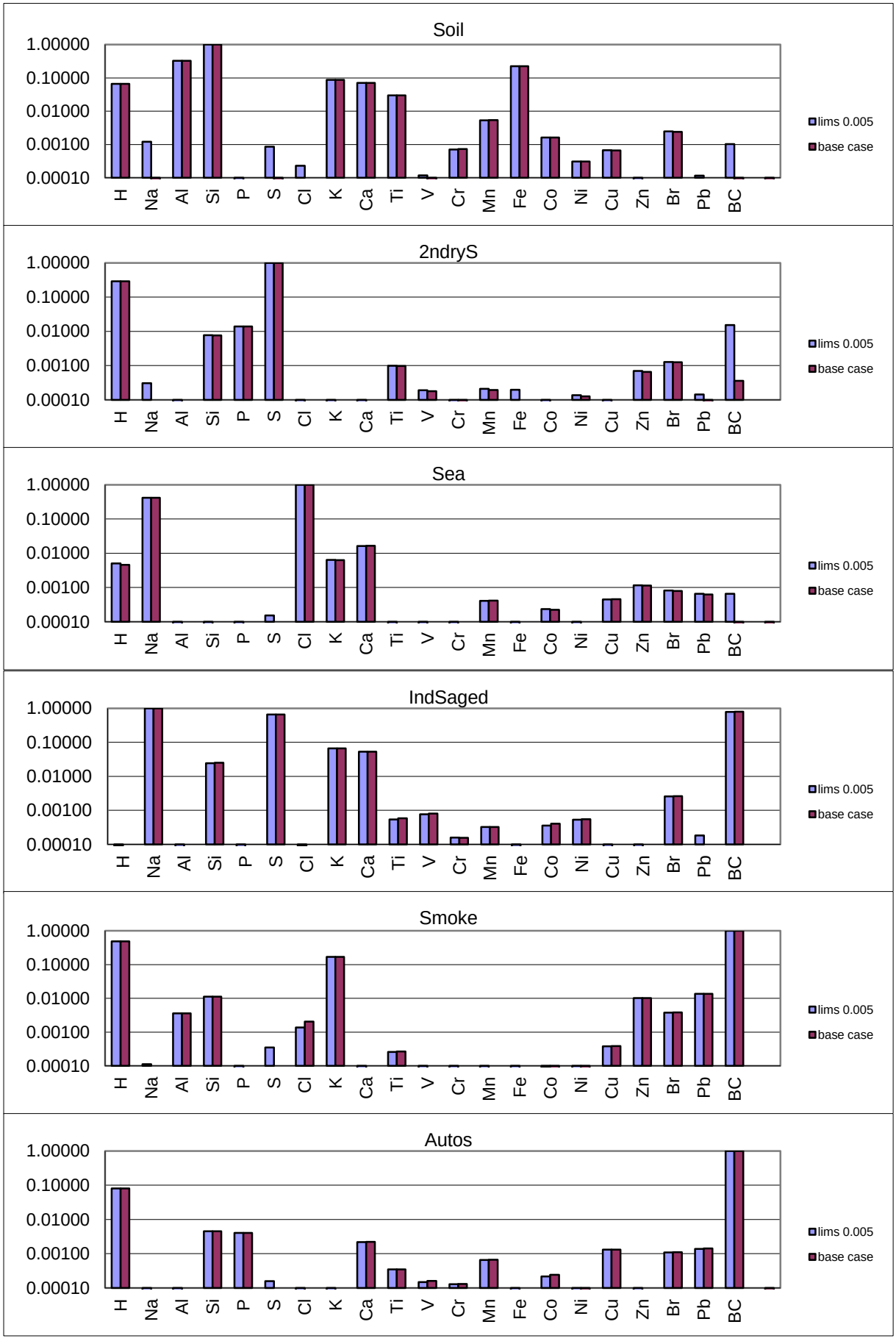
Level 1 χ^2			Level 2 χ^2			Level 3 χ^2		
Value	Iterations	Q_{expt}	Value	Iterations	Q_{expt}	Value	Iterations	Q_{expt}
5.000	158	6708	1.00	163	6699	1.00	158	6710
0.500	174	6709	0.010	174	6709	0.010	169	6709
0.050	191	6709	0.0010	180	6709	0.00010	174	6709
0.005	204	6708	0.0001	180	6709	0.000001	174	6709

Table 8: The S/H ratio for *2ndyS* fingerprint, the Al/Si ratio for the *Soil* fingerprint and the Cl/Na ratio for the *Sea* fingerprint for the various values of the χ^2 parameter, compare to S/H ratio of 3.46, Al/Si ratio of 0.325 and Cl/Na ratio of 2.38 for the *base case*.

Level 1 χ^2				level 2 χ^2				Level 3 χ^2			
Value	S/H	Al/Si	Cl/Na	Value	S/H	Al/Si	Cl/Na	Value	S/H	Al/Si	Cl/Na
5.000	3.46	0.325	2.38	1.0000	3.44	0.325	2.39	1.0000	3.41	0.325	2.38
0.050	3.46	0.325	2.38	0.0010	3.46	0.325	2.38	0.0100	3.45	0.325	2.38
0.005	3.46	0.325	2.38	0.0001	3.46	0.325	2.38	0.0001	3.46	0.325	2.28

Table 9: Percentage contribution to the total mas from each of the identified fingerprints, when the different χ^2 values are used.

	base case	Level 1 χ^2			Level 2 χ^2			Level 3 χ^2		
χ^2value		5.000	0.050	0.005	1.00	0.0010	0.0001	1.00	0.010	0.0001
2ndryS	25.18	25.17	25.23	25.17	25.14	25.18	25.18	25.47	25.28	25.18
Soil	5.59	5.59	5.59	5.59	5.59	5.59	5.59	5.60	5.59	5.59
Sea	5.17	5.18	5.16	5.18	5.15	5.17	5.17	5.12	5.14	5.17
IndSaged	14.46	14.46	14.46	14.47	14.49	14.46	14.46	14.36	14.43	14.46
Smoke	33.77	33.75	33.79	33.75	33.74	33.75	33.75	33.60	33.80	33.77
Autos	14.51	14.52	14.47	14.52	14.52	14.52	14.52	14.40	14.46	14.51
Ind.	1.32	1.33	1.30	1.32	1.37	1.33	1.33	1.45	1.30	1.32


Soil
2ndryS
Sea
IndSaged
Smoke
Autos
1.00000
0.10000
0.01000
0.00100
0.00010
H Na Al Si P S Cl K Ca Ti V Cr Mn Fe Co Ni Cu Zn Br Pb BC
lims 0.005
base case

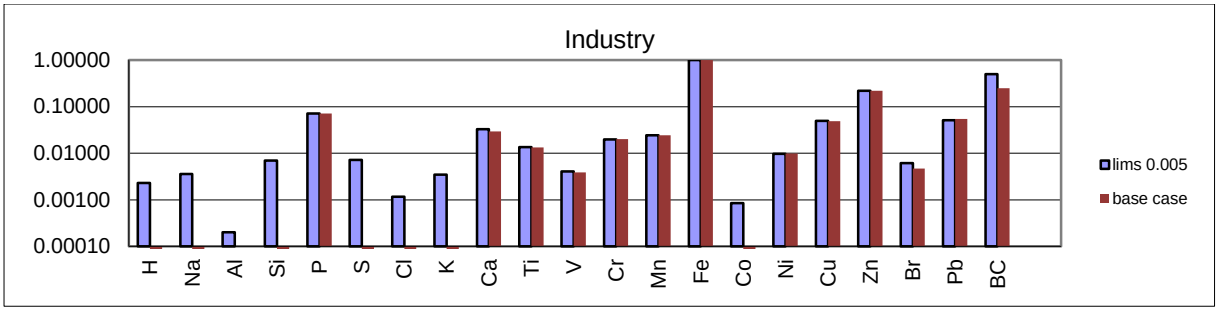


Figure 8: Fingerprint comparison for *base case* (brown) and when the level 3 *lims* value is set to 0.005 (blue).

8 Rotational freedom using *Fpeak*

The rotational freedom for the *base case* was examined by changing the *Fpeak* parameter by 0.1 both in the negative and positive direction from 0. Little difference in the factor composition could be seen for *Fpeak* values between -0.5 and 0.2. When an *Fpeak* of 0.3 was used a number of the factors could not be attributed to possible source fingerprints and when an Fpeak of -0.6 was used S and BC appeared in the *Soil* fingerprint and the *Industry* fingerprint had a high proportion of H and K. The smallest Q_{expt} value was obtained when *Fpeak* of zero was used (Table 10, Figure 9: indicating that values of *Fpeak* between -0.4 and 0.2 are acceptable, as the Q_{expt} value changes little in this region), and the corresponding contributions to the total mass from each identified fingerprint are presented in Table 11. While some differences in the BC component of the F matrix factors for *2ndryS* and *Industry* were seen there was little difference in the percentage contribution by each of the identified fingerprints. However, there were large differences in the factors when *Fpeak* values below -0.5 and above 0.2 were used.

Table 10: Statistics for different values of the *Fpeak* parameter.

Value	Iterations	Q_{expt}	S/H	Al/Si	Cl/Na
-0.6	340	7003	3.02	0.32	1.85
-0.5	331	6925	3.05	0.324	1.92
-0.4	288	6768	3.17	0.326	2.08
-0.3	180	6733	3.27	0.325	2.21
-0.2	188	6718	3.31	0.325	2.29
-0.1	190	6710	3.39	0.325	2.35
0.0	174	6709	3.46	0.325	2.38
0.1	183	6715	3.49	0.325	2.54
0.2	447	6728	3.54	0.325	3.01
0.3	244	7298	0.16	0.325	1.44

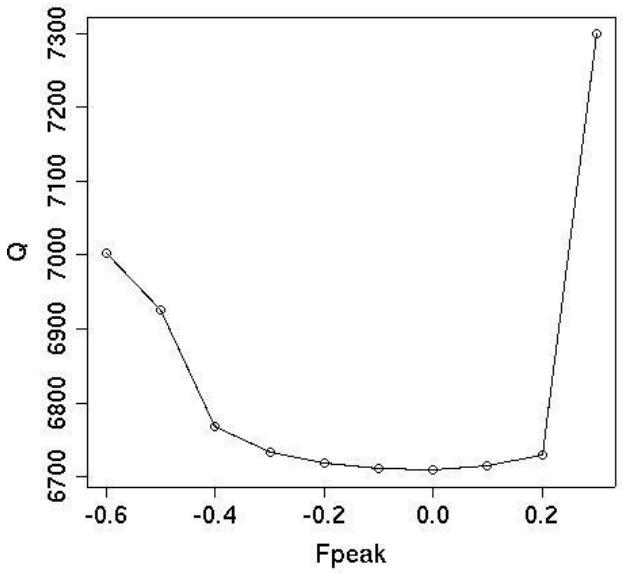


Figure 9: The Q_{expt} value with varying *Fpeak*.

Table 11: Percentage contribution by each source type to the total $PM_{2.5}$ mass for each of the *Fpeak* values.

Fpeak	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2
2ndryS	25.62	25.76	25.79	25.51	25.18	25.24	25.34
Soil	5.45	5.50	5.54	5.58	5.59	5.61	5.67
Sea	5.84	5.40	5.27	5.20	5.17	5.04	4.74
IndSaged	13.97	14.23	14.28	14.38	14.46	14.41	14.26
Smoke	33.35	34.01	34.09	33.99	33.77	33.56	33.22
Autos	13.69	13.28	13.51	13.98	14.51	14.92	15.49
Ind.	2.08	1.82	1.52	1.36	1.32	1.22	1.28

9 Non zero values for below detection limit values

Values for Na which were below the detection limit were replaced by MDL/3 and then by MDL/2. Little difference was seen for MDL/3, but some small differences were seen for MDL/2 (Table 12). When replacing below detection limit values of Na by MDL, Na appeared in the *Soil* and *2ndryS* fingerprints (Figure 10). The Q_{expt} values were 6622, 6597, 6578 when MDL/3, MDL/2 and MDL were used, respectively, compared to the *base case* Q_{expt} value of 6709. The ratio R_Q, where Q_{thry} is calculated according to equation 5 (Q_{thry}=12621) were as follows, 0.53, 0.52, 0.52 and 0.52 for the *base case* and when measurements with zero values were replaced by MDL/3, MDL/2 and MDL, respectively.

Table 12: Percentage contribution to the total mass from each of the identified fingerprints when the zero values of Na were replaced by MDL/3, MDL/2 and MDL, in turn. Also included are results for the next section, i.e. when the measurement uncertainty is varied.

	base case	MDL/3	MDL/2	MDL	$0.5*s_{i,j}$	$0.1*s_{i,j}$	$2*s_{i,j}$	$5*s_{i,j}$
2ndryS	25.18	24.87	24.77	24.67	25.86	21.10	25.60	29.56
Soil	5.59	5.59	5.59	5.59	5.50	5.89	5.60	6.43
Sea	5.17	5.18	5.23	5.32	4.95	6.51	5.03	3.71
IndSaged	14.46	14.87	15.05	15.34	14.13	27.23	14.13	9.92
Smoke	33.77	33.12	32.85	32.14	34.19	26.64	33.63	41.57
Autos	14.51	14.83	14.93	15.20	14.17	9.87	14.53	6.81
Ind.	1.32	1.54	1.58	1.74	1.20	2.76	1.48	2.00

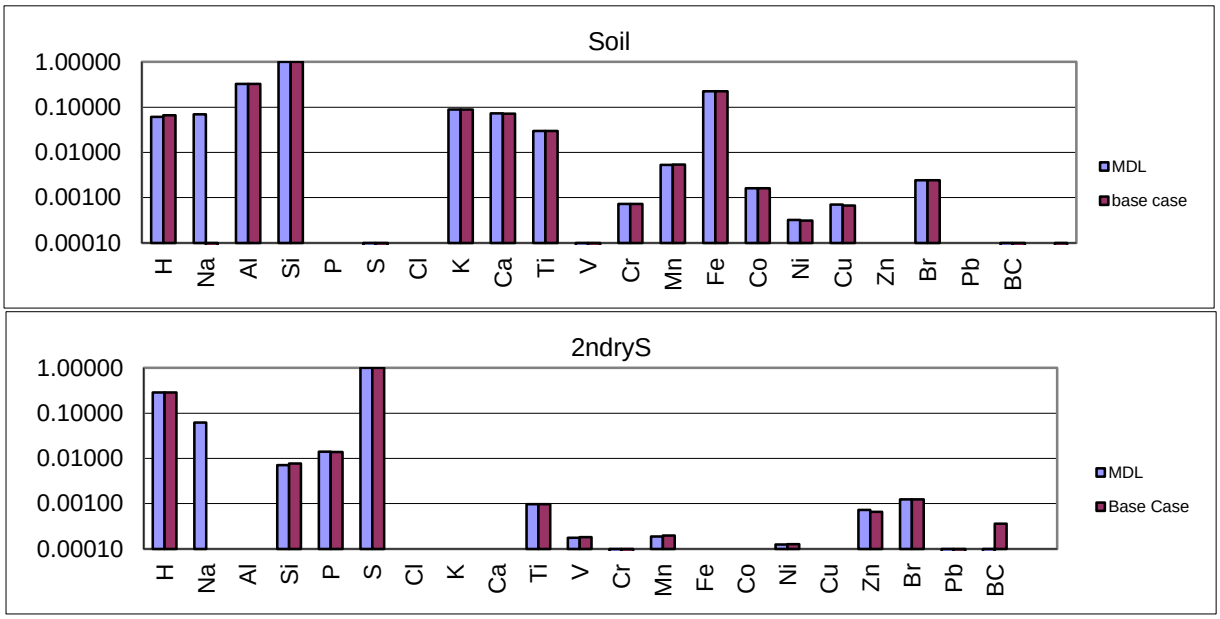


Figure 10: Difference in *Soil* and *2ndryS* factor between the *base case* (brown) and when below detection limit Na values are replaced by the MDL (blue).

10 Varying the measurement uncertainty

For the optimisation as specified in equation 3 an uncertainty, $s_{i,j}$, is associated with each element for each sample. This uncertainty is a combination of the MDL and statistical error for each measurement. Here we apply the same scaling to both the MDL and the error to gauge the impact the uncertainty has on the final solution. When the measurement uncertainty was halved some small differences in percentage contributions by fingerprints were seen to those for the *base case*, as seen in Table 12. When the measurement uncertainty was doubled little difference was seen. The main differences were when a 10^{th} of the measurement uncertainty was used and mainly in the *Smoke*, *Autos* and *Industry* fingerprints (Figure 11 and Table 12). On the other hand when the uncertainty was multiplied by 5 a factor containing mostly Cl and BC appeared, indicating that problems are likely when the uncertainties are too big, as well as when they are too small. The ratio of key elements for some of the fingerprints is presented in Table 13, and so are the number of iterations to convergence and the

final Q_{expt} value. A ratio of more than 1 of R_Q gives an indication of the specified uncertainties might have been too small. Similarly a very small ratio might indicate that the specified uncertainties are too high. In addition to the Q_{expt} value the scaled residuals for each element can be examined (Figure 12). These should be normally distributed with a mean of zero. Large values for some elements might indicate that the uncertainties for that element might have been set too small.

Table 13: The ratios of key elements in the *2ndryS*, *Soil* and *Sea* fingerprints, when the measurement error is halved and a 10th of the measurement error is used.

	base case	$0.5*s_{i,j}$	$0.1*s_{i,j}$	$2*s_{i,j}$	$5*s_{i,j}$
S/H	3.46	3.40	0.26	3.58	3.77
Al/Si	0.325	0.325	0.321	0.325	0.319
Cl/Na	2.38	2.50	1.38	2.49	_
Iterations	174	184	227	376	264
Q_{expt}	6709	22612	211192	1831	934
R_Q	0.53	1.79	16.73	0.15	0.07

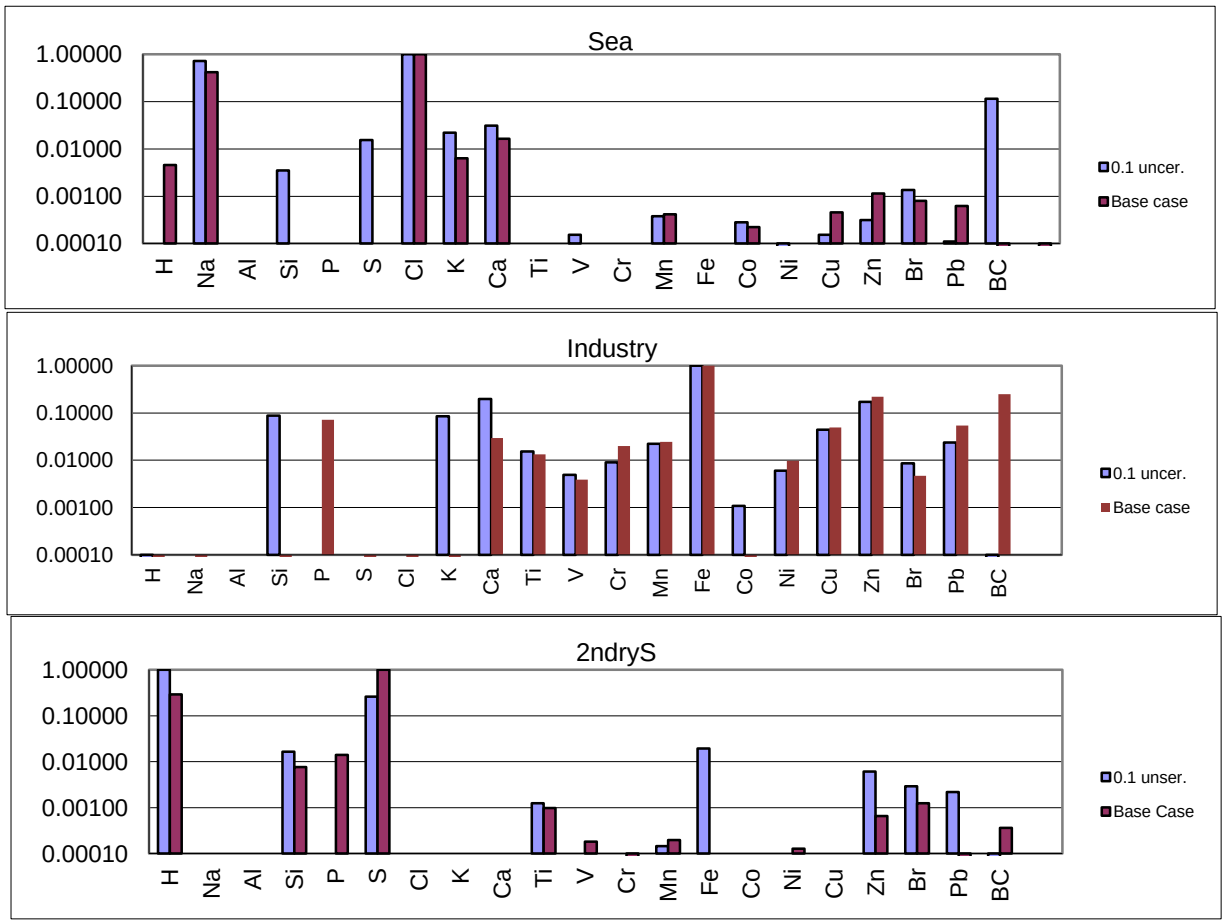


Figure 11: Difference in selected fingerprints between the *base case* (brown) and when the uncertainty is replaced by a 10th of the uncertainty (blue).

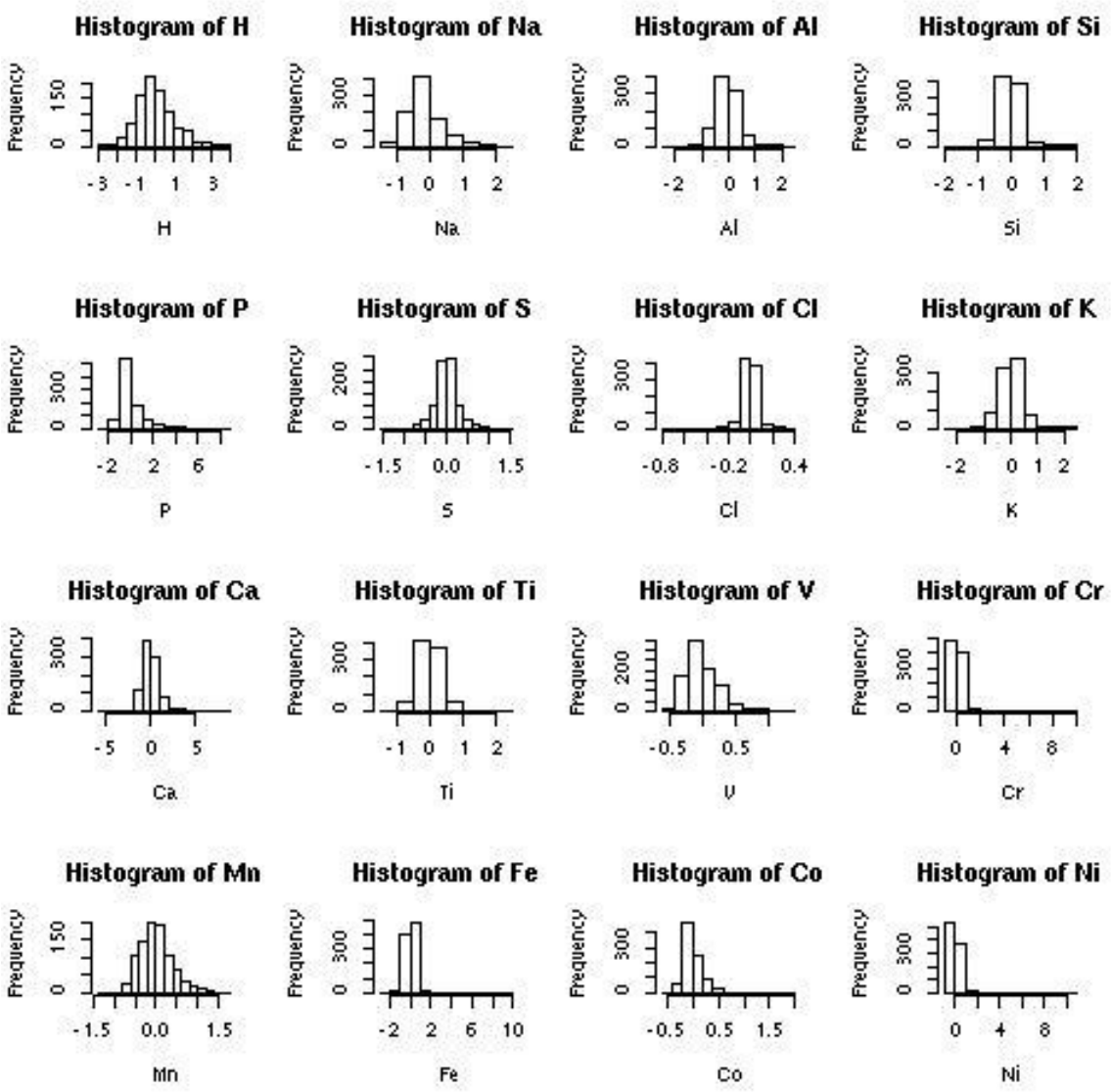


Figure 12: Scaled residuals for selected elements of the *base case*.

11 Adding the total mass to the PMF analysis

Introducing the total mass into the analysis was achieved by adding the total mass to the first column of the data matrix and then also adding the corresponding MDL and error to the first column of the mdl and error matrix, respectively. For the solution, the PMF2 MaxF=1 option was set to True. The solution resulted with a value of one for the entry corresponding to the total mass for each factor of the F matrix; hence the columns of the G matrix correspond to the amount of the total sample mass apportioned to each of the recovered fingerprints. Thus, from the G matrix one can directly calculate the percentage contribution to total mass from each of the identified source fingerprints.

When the total mass is not used in the PMF2 analysis, the way of estimating the percentage contribution to the total mass from each of the identified fingerprints is to carry out a multi-linear lest squares using the G matrix as the independent variables and using the total mass as the dependent variable. This process gives a normalisation factor for each of the fingerprints.

A multi-linear regression was carried out on the G matrix obtained when the total mass was added to the PMF2 analysis, to compare the source apportionment based on the two possible approaches. The coefficients of the multi-linear fit, which are close to one, are presented in Table 14 and the corresponding percentage contributions in Table 15.

Table 14: The coefficients of the multi-linear least squares fit to the G matrix generated by PMF when total mass was added to the analysis.

	$s_{i,j}$= 4	$s_{i,j}$=1	$s_{i,j}$=0.1
2ndryS	1.02	1.02	1.01
Soil	1.14	1.14	1.12
Sea	1.18	1.18	1.11
IndSaged	1.12	1.12	1.14
Smoke	1.05	1.05	1.03
Autos	1.30	1.30	1.25
Ind.	0.82	0.81	0.80

Adding the total mass to the PMF analysis resulted in little change for error (on the mass) = 4, 1, and 0.1. When the error on the mass was reduced further more changes were seen. The S/H ratio of 3.47, 3.47 and 3.43 was seen for error=4, 1, 0.1, respectively. The Al/Si ratio remained unchanged. The Cl/Na ratio was 2.42, 2.41 and 2.35 for error=4, 1, 0.1, respectively, compared to 2.38 for the *base case*.

Table 15: Percentage contribution from each fingerprint to the total mass when different errors were specified for the measured total mass, using the G matrix and when using a multi-linear least squares fit in addition to the PMF analysis.

		Multi-linear least squares			From G matrix		
	base case	$s_{i,j}$= 4	$s_{i,j}$=1	$s_{i,j}$=0.1	$s_{i,j}$= 4	$s_{i,j}$=1	$s_{i,j}$=0.1
2ndryS	25.18	25.22	25.23	25.46	26.98	26.96	26.90
Soil	5.59	5.58	5.58	5.52	5.31	5.32	5.31
Sea	5.17	5.15	5.15	5.23	4.74	4.76	5.04
IndSaged	14.46	14.45	14.45	14.32	13.99	13.96	13.49
Smoke	33.77	33.72	33.71	33.79	35.07	35.07	35.16
Autos	14.51	14.55	14.55	14.38	12.13	12.14	12.35
Ind	1.32	1.33	1.33	1.30	1.78	1.79	1.75

12 Effects when different starting guesses are used (different seeds)

For this dataset there was little difference in the results when different seeds were used.

Table 16: The ratios of key elements in *2ndryS*, *Soil* and fresh sea spry (*Sea*) when different seeds were used. Also included are ratios when H is pulled down to zero from *Soil* in a restart run and when the uncertainty for BC is increased (see next two sections)

	base case	Seed=1	Seed=100	Restart *base case*	BC $2*s_{i,j}$	BC $5*s_{i,j}$
S/H	3.46	3.46	3.46	3.44	6.04	5.89
Al/Si	0.325	0.325	0.325	348	0.326	0.327
Cl/Na	2.38	2.38	2.38	2.47	5.03	5.04
Iterations	174	175	151	162	167	167
Q_{expt}	6709.04	6708.81	6708.98	6730.47	5758.77	4865.89
R_Q	0.53	0.53	0.53	0.53	0.46	0.39

Table 17: Fingerprint percentage contribution to the total mass when different seeds were used. Also included are the percentage contributions when the H was pulled to zero from *Soil* in a restart run and when the BC uncertainty is increased (see next two sections).

	base case	Seed=2	Seed=100	Restart *base case*	BC $2*s_{i,j}$	BC $5*s_{i,j}$
2ndryS	25.18	25.17	25.17	25.28	26.92	27.66
Soil	5.59	5.59	5.59	4.96	5.48	5.49
Sea	5.17	5.15	5.16	5.29	3.87	3.41
IndSaged	14.46	14.46	14.46	14.57	9.03	8.48
Smoke	33.77	33.72	33.77	33.41	23.66	23.6
Autos	14.51	14.56	14.53	15.24	26.43	26.85
Ind	1.32	1.35	1.32	1.25	4.61	4.51

13 Pulling down of factor elements

For the *base case* solution, the *Soil* fingerprint contained H (Figure 6). A solution was obtained by pulling down to zero the H in *Soil* fingerprint by using the FKey option and restarting from the good solution. Following the restart, with pulling down to zero only the H in *Soil*, the main differences were in the *Soil*, *Sea* and *Industry* fingerprints (Figure 13). The Q_{expt} value increased slightly, i.e. from 6709 to 6730. The ratios of key elements are presented in Table 16 and the percentage contribution from each fingerprint to the total mass is given in Table 17.

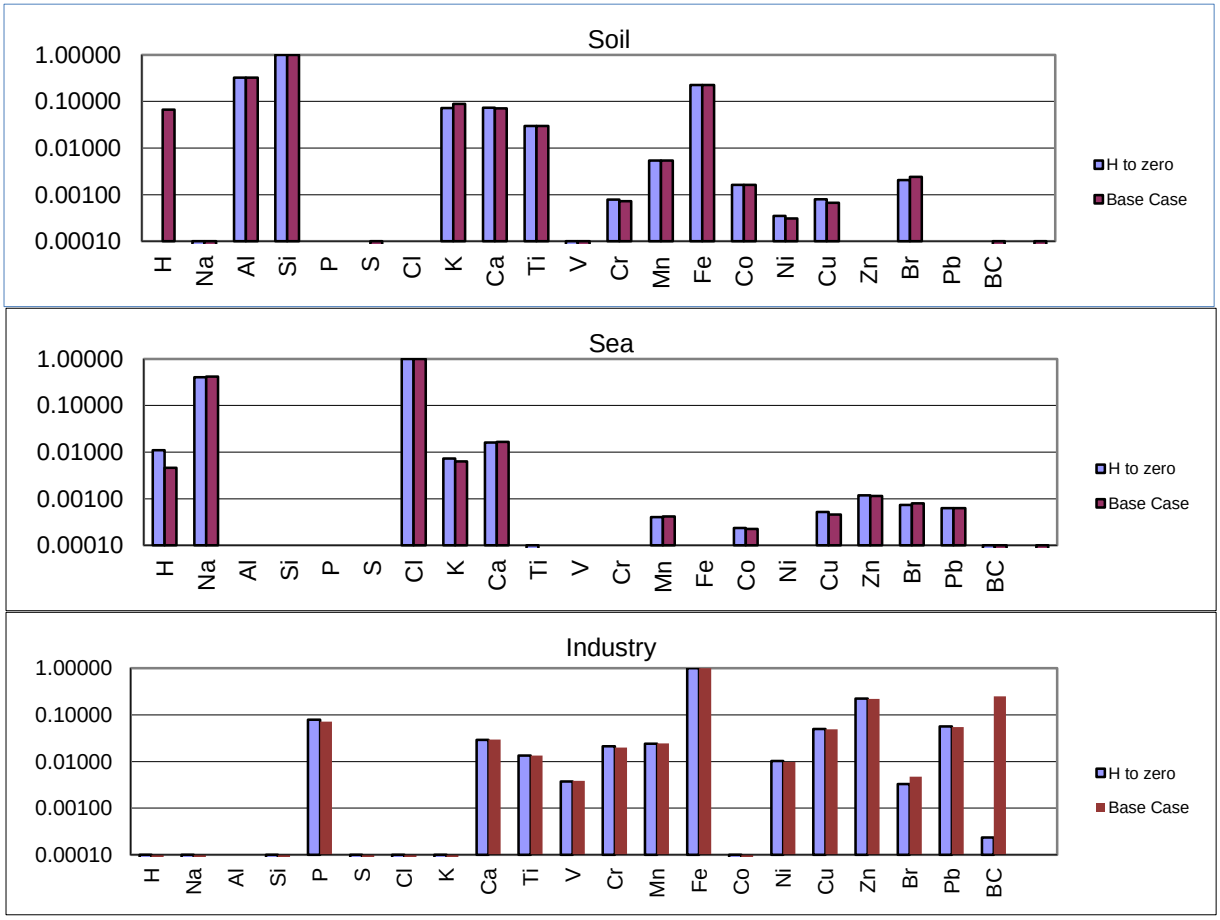
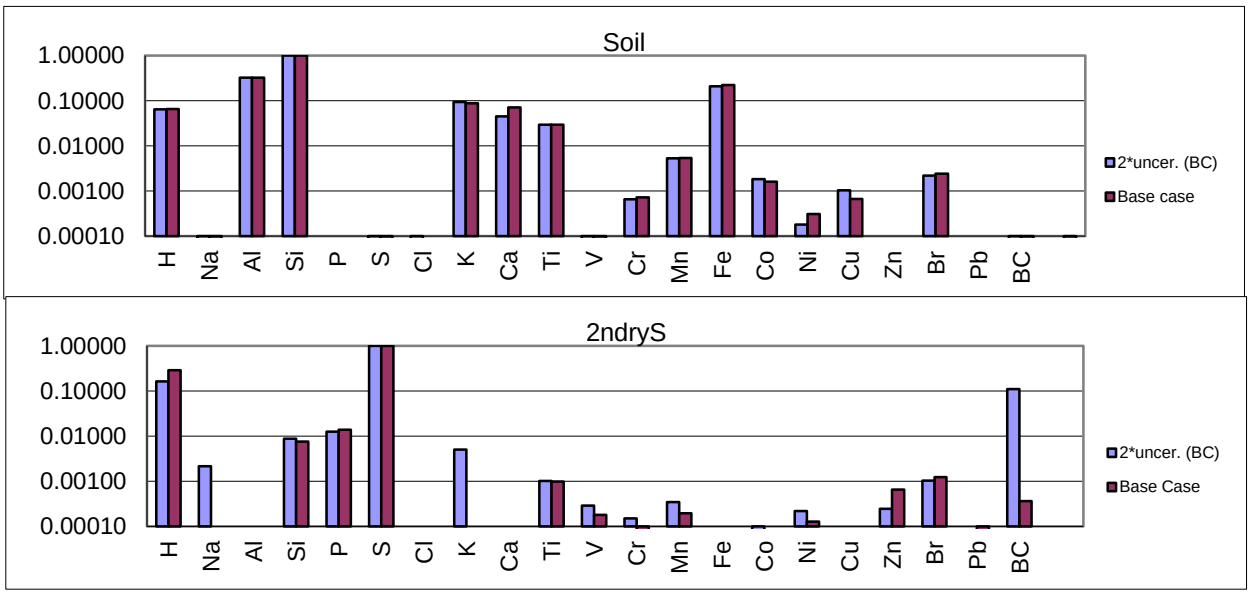


Figure 13: Differences in the *Soil, Sea* and *Industry* fingerprints after starting from a good solution and pulling down to zero the H in *Soil*.

14 Uncertainty in Black Carbon

In this dataset the largest concentration occurs for BC and BC is present in a number of fingerprints. BC also has a small measurement error. Here we investigate the behaviour of BC as we increase its $s_{i,j}$ by a factor of 2 and a factor of 5, while all the $s_{i,j}$ of all other elements are kept to that of the *base case*. Considerable difference is seen between the *base case* and the two cases when the uncertainty of BC is increased (Table 16, Table 17 and **Figure 14**), while the difference between the two cases with larger uncertainty on BC is small. For example, BC and Na appeared in the *2ndryS* fingerprint, Fe appeared in *IndSaged* while the Fe fraction was less in the *Industry* fingerprint.



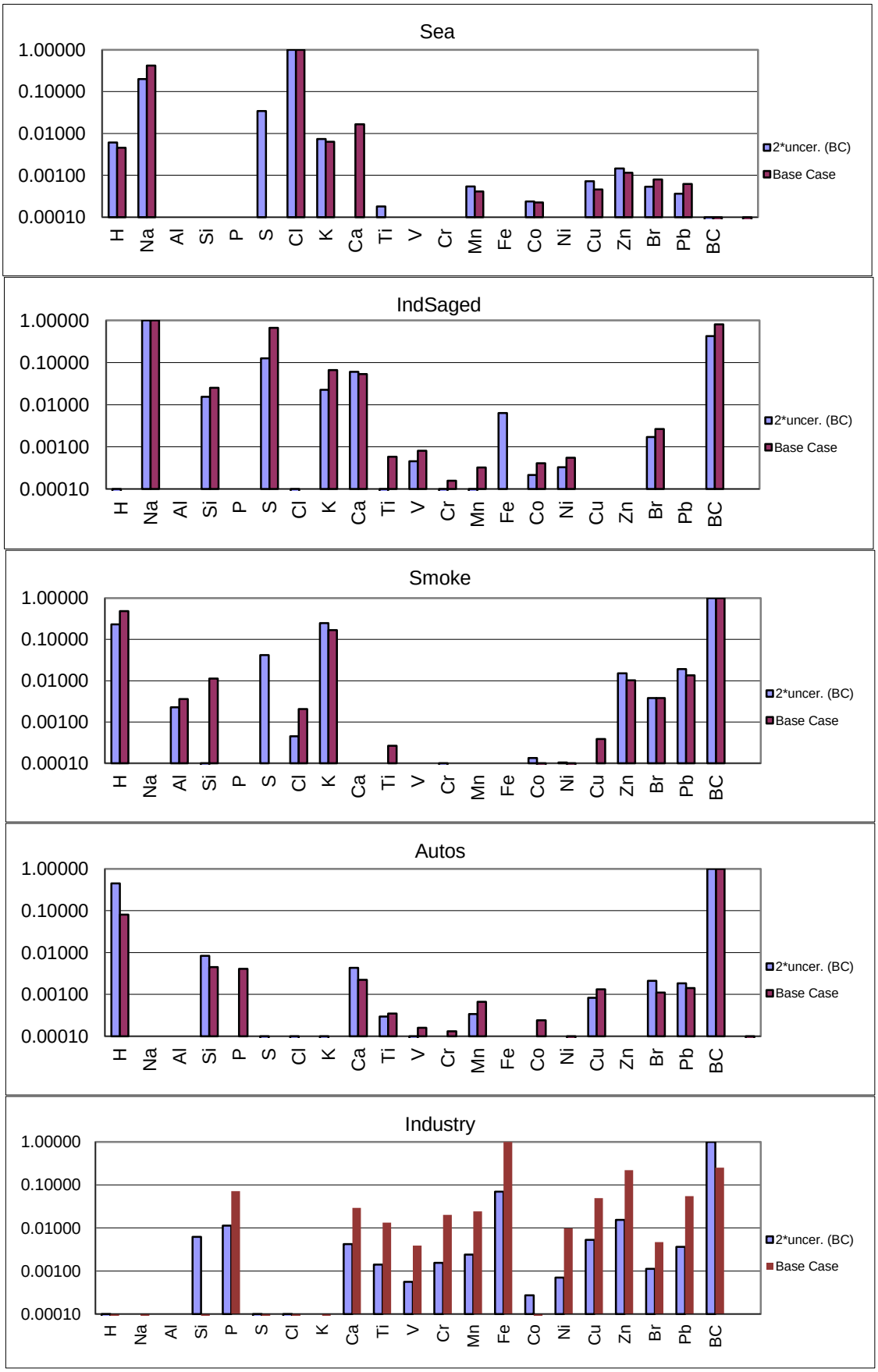


Figure 14: Fingerprints for the *base case* (brown) and those when the BC uncertainty ($s_{i,j}$) was doubled.

15 Pre-processing of the data

The measurement of the elemental species allows estimates of significant fine particle components like ammonium sulfate, soil and sea salt to be made according to Malm et al., (1994):

$$\text{Salt} = 2.54[\text{Na}] \quad (6)$$

$$\text{Ammonium Sulfate} = 4.125[\text{S}] \quad (7)$$

$$\text{Soil} = 2.20[\text{Al}] + 2.49[\text{Si}] + 1.63[\text{Ca}] + 1.94[\text{Ti}] + 2.42[\text{Fe}] \quad (8)$$

$$\text{Smoke} = [\text{K}] - 0.6[\text{Fe}] \quad (9)$$

$$\text{Organics} = 11[\text{H}] - 0.25[\text{S}] \quad (10)$$

$$\text{RCM} = \text{Salt} + \text{Ammonium Sulfate} + \text{Soil} + \text{Smoke} + \text{Organics} + \text{BC} \quad (11)$$

where the [ ] represent the concentration of that element. Equ. (7) assumes all the sulfate is fully neutralised and occurs on the filter as ammonium sulfate. The organic matter estimate, equ. (10), assumes the average $PM_{2.5}$ organic composition was 9%H, 20%O and 71%C (Malm et al., 1994).

The Reconstructed Mass (RCM), as defined in equ. (11) can be compared with the gravimetric mass for each filter to determine the degree of mass closure. High mass closure is essential for good, reliable source apportionment of the total measured mass. Thus before a PMF analysis is carried out, plotting the samples RCM against the gravimetric mass and then removing any outliers from the analysis will ensure that the good samples are used in the PMF analysis. A rule of thumb is to remove samples that are outside between three and four standard deviations from the regression line of best fit.

16 Conclusions

For this dataset the number of factors chosen to be recovered and the errors in the data were the key quantities which affect the recovered fingerprint. Care needs to be taken when specifying the MDLs and the errors on the measurements. Problems arise when these are too small and also when they are unnecessarily high.
Specifically:

- The *lims* options mainly affected the number of iterations to convergence and how closely trace elements of the F matrix were pulled to zero.
- Care needs to be taken when using *Fpeak* to ensure that Q_{expt} does not increase significantly and ratios of key elements can also be used to identify the extent of the rotation.
- Replacing zero measurements by fractions of MDL affects the solution; however, our preference is to alter the uncertainties.
- Varying the uncertainties of the measurement has an impact on the solution and the R_Q ratio can be used as an indicator of possible problems with the specified uncertainties, i.e. a ratio significantly greater than one may indicate that the uncertainties have been set too small and a small ratio may indicate that the uncertainties have been set too high.

17 One possible approach for analysis

Once the data and the uncertainties have been prepared the following procedure could be used:

- Estimate the RCM as in section 15 and remove any outliers.
- Ensure that the MDLs and errors are reasonable. Errors may be 5% to 100%.
- Obtain a solution using 6 to 10 factors
- Identify any outliers, by plotting the PMF fitted mass against the measured mass, and remove any outliers from the dataset.
- Select the number of fingerprints
    - Examine the $R_Q=Q_{expt}/Q_{thry}$, if greater than one then more fingerprints might be needed on the other hand if much less than 0.5 than less fingerprints might be needed.
    - However, R_Q can also be affected by the uncertainties, if R_Q is much larger than one the uncertainties may be too small, or if R_Q is much less than 0.5 then the uncertainties may be too high.
    - Examine the fingerprint, can they be attributed to known sources.
    - Examine correlation between the columns of the G matrix. Correlation might indicate that a single source has been split into two fingerprints.
- Select possible number of fingerprints from above steps.
- Pull down BC and H from the Soil fingerprint if required.
- Examine R_Q and scaled residuals and ratios of key elements (Al/Si in soil, S/H in secondary sulfur and Cl/Na in sea salt) to gauge if the uncertainties are reasonable.
- Use *Fpeak* to examine rotational freedom and check its influence on known elemental ratios like Al/Si in soil and Cl/Na in sea spray etc.

18 Acknowledgements

The authors would like to thank Ed Stelcer for reviewing this report.

19 References

Cohen D.D., Bailey G.M. and Kondepudi R., 1996: Elemental analysis by PIXE and other IBA techniques and their application to source fingerprinting of atmospheric fine particle pollution, Nucl. Instr. and Methods 109, 218-226.

Cohen D.D., 1998: Characterisation of atmospheric fine particles using IBA techniques. Nuclear Instruments and Methods in Physics Research B136-B138, 14-22.

Cohen D.D., Taha G., Stelcer E., Garton D., Box G., 2000: The Measurement and Sources of Fine Particle Elemental Carbon at Several Key Sites in NSW over the Past Eight Years. 15th International Clean Air Conference, Sydney Australia.

Cohen D.D, Garton D., Stelcer E., Hawes O., 2004: Accelerator based studies of atmospheric pollution processes. Radiation Physics and Chemistry 71, 758-567.

Cohen D.D., Crawford J., Stelcer E., Atanacio A.J., 2012. Application of positive matrix factorization, multi-linear engine and back trajectory techniques to the quantification of coal-fired power station pollution in metropolitan Sydney. Atmospheric Environment, 61, 204-211.

Henry R.C., 2002. Multivariate receptor models – current practice and future trends. Chemometrics and Intelligent Laboratory Systems, 60, 43-48.

Hopke P.K., 1991. Receptor Modeling for Air Quality Management. Elsevier, Amsterdam.

Jollife I.T., 1986: Principal Component Analysis. Springer, New York.

Malm, W.C., Sisler, J.F., Huffman, D., Eldred, R.A., Cahill, T.A., 1994. Spatial and seasonal trends in particle concentrations and optical estimations in the US. J. Geophys. Res. 99, 1347-1370.

Paatero, P., Tapper, U., 1994: Positive Matrix Factorisation: A non-negative factor model with optimal utilisation of error estimates of data values. Environmetrics, 5, 111-126.

Paatero P., 2010. User's Guide for Positive Matrix Factorisation programs PMF2 and PMF3, Parts 1 and 2.

Seinfeld J.H., Pandis S.N., 1998. Atmospheric Chemistry and Physics. From Air Pollution to Climate Change. John Wiley and Sons, Inc., USA, ISBN 0-471-1-17816-0.

Taha G., Box G.P., Cohen D.D, Stelcer E., 2007: Black Carbon Measurement using Laser Integrating Plate Method. Aerosol Science and Technology, 41, 266-276.